4/24/63

October 2, 1963

Dr. James F. Crow
The University of Wisconsin
Department of Medical Genetics
Genetics Building
Madison 6, Wisconsin

Dear Jim:

I was delighted to have the chance to talk to you today.

Probably I would not have reacted so quickly to the census project suggestion that the NIH brought up were it not for the difficulties that I seemed to be running into in my earlier efforts to get some very modest work going in this direction. I would not have had the courage to go in with a proposal of this magnitude myself, and also doubt whether it would have obtained much support without this connection. In any case I think the opportunity to get into the 5% census sample is a unique one, and the Institute's cooperation in dealing with the Census Bureau means a good deal more than just in obtaining the necessary funding.

Enclosed is a listing of the characteristics available on the 1 per 1000 sample tapes, and you will also note the items that we have been planning to pull off for our own brief summary tape to allow more efficient recurrent use of the same file. A good deal of the information capacity of the file is wasted from our point of view, since it comprises so much marking information and also since a great deal of it is repeated for every member of the household and need be stated only once. We have therefore programmed (and are waiting for a completely successful run) a copy with variable length blocks corresponding to the number of individuals in each household. An early reaction I have to the largerfile is that it probably would be most justifiable to adopt the same approach since tape-passing time will generate a large part of the total computational costs, and at worst it can cost only about an additional third to make a full tape pass once to genetate the summary copy, and then tabulate from the latter. This will also allow the flexibility of a much cheaper access for further runs if these are indicated. I should have mentioned that the costs of processing these ten million or so records will be of the order of $100,000. Another thing that should probbbly be done in the first pass is to re-sort the file into about ten categories by some relevant classification which will permit a more reasoned search of each sub-file. Size of household unit, or rather number of offspring present, might be the aptest variable since one can write much simpler programs for dealing with the very small families as compared to the large ones, and it may be quite à propos to deal with a smaller sample of the more prevalent families than of the rarer ones.

Although some questions about family structure are what led me into this game, I would put first priority as a matter of general scientific interest on the questions of differential fertility, along the lines that you have raised, and I think this inquiry will be a unique opportunity to make an adequate analysis of the variance of fertility with joint distributions of a number of factors which are presently unavailable. Before we make a final commitment on the specifications for this run, I hope we will have an opportunity to see the tabulations that have been constructed or are planned for with respect to the distribution of offspring from the 1960 census along the lines of the 1950 Special Reports, Volume IV. There will be no point in repeating other tabulations already in progress, although I suppose we can be reasonably confident that there will not be much breakdown giving detail on specified numbers of children. In any case, as you mentioned, the lack of details on the tales, e.g. that such categories as "white, laborers, women 45 to 49' will be represented only as 23.4%: 5 or more--well, this is no help at all in a precise analysis. I suppose one thing that ought to be computed for each family is a net replacement index, something along the lines of the expectation of numbers of offspring after a fixed time interval, assuming that each of the progeny had the same reproductive pattern with respect to total fertility and maternal age at each birth as did the parents. Such a figure would be somewhat more precise than the ones you have used so far without age correction, and could have the additional advantage of incorporating more recent information on women who may not have completed their fertility.

I am a little puzzled how we can hope to handle as many tabulations as we will soon find we would like to have. I don't yet see how we can spread the over more cells than are available in the fast memory of the computer, for which something like 30,000 words may be a rather optimistic estimate. It just wont't take very many cross tabulations to saturate this kind of memory. We are, however, looking into a number of ways of using split-word or single bit formats that might give us back more in useful memory capacity than it costs in additional computer time. Here is precisely where the pilot runs on the 1 per 1000 sample tape will be invaluable.

I haven't begun to think how to handle the problem of representing family structures in a way that preserves the most essential information and yet does not require storage for an inordinately large number of possible combinations, e.g. in terms of the respective intervals between successive births. This is going to perhaps the most thought, including the question how best to display the results of any tabulations that do become available.

I am also enclosing some of the summary data from the 1 per 1000 sample, which should be quite representative of the 5% sample and need to be multiplied accordingly.

Looking forward to seeing you next month, and I hope you do find it possible to follow through on this.

Sincerely,

Joshua Lederberg
Professor of Genetics